7.344 Genomics and bioinformatics of gene expression

Assignment 2:
Cross-species analysis of mucle-specific promoter regions

General guidelines

1. Length about 2 pages

2. Due: Wed. 11/19/2003 at the beginning of the class or by email.

3. Do not copy and paste from papers. It s boring for you and for us. Be specific without getting lost in the details.

Specific guidelines

Go to the ENSEMBL database of eukaryotic genomes, www.ensembl.org. Search for "alpha-skeletal actin", a muscle-specific gene, and take the best hit to the "homo sapiens" genes. Make sure you really got what you wanted, and look at the information in the report for this gene. Then go to "export data", and retrieve the sequence in FASTA format with a context of 500 bp. Save the *last* 600 or so bases as the potential promoter region (note: the gene is on the reverse strand!).

Go back to the gene report and from there to the homologous mouse gene. Retrieve the mouse promoter region as above. Then go to www.wadsworth.org/resnres/bioinfo/ and from there to the Bayesian Phylogenetic Footprint (that is the alignment program in the paper by Wasserman et al, "Human-mouse genome comparsions to locate regulatory sites"). Align the two sequences.

Now, go back to the human gene report on ENSEMBL and to the zebrafish homolog (danio rerio). Retrieve the promoter region, but this time take a context of 1,900 bases. Again, save the last 600 bases. Then go back to the Bayesian Phylogenetic Footprint and align the human against the fish sequence.

1. What is the length of the transcript for the human, mouse, and fish ortholog? How many exons are annotated in each species? Why did you have to use a context of 1,900 bases upstream for the fish gene

instead of 500? (hint: look at the graphical gene structure of the genes in the ENSEMBL reports).

2. What can you see in the human-mouse alignment? Can you tell potential muscle-specific or other transcription factor binding sites? (refer to the two papers by W. Wasserman). If so, which ones? Remember, you are looking at the *reverse complement* of the promoter region, as the genes were both located on the reverse strand (e.g. for AAACCTG the reverse complement is CAGGTTT).

3. What is different in the human-fish alignment? Are there any conserved regions in common with the human-mouse alignment? Do they match known binding sites? Which regions are not in common?

4. Imagine you did not know anything about muscle-specific binding sites, but that you had a set of about 20 co-regulated genes from human, mouse and fish. Describe *one* approach we discussed to identify potential transcription factor binding sites, and which you believe could actually be useful for vertebrates. You can, but do not have to, use the sequence from one or both of the related organisms.

Don't forget to include the alignments you obtained in your reports.